Solutions of the exercises on Linear Regression and Analysis of Variance (part 3 Applied Statistics)

Exercise 1
a. \( y = \beta_0 + \beta_1 x + \epsilon \), where \( y \) = “yearly amount spend on food” and \( x \) = “the yearly income” and the independent disturbances \( \epsilon \) all have a \( N(0, \sigma^2) \)-distribution (the parameters \( \beta_0 \), \( \beta_1 \) and \( \sigma^2 \) are unknown parameters

b. Using the coefficients table of the output: \( \hat{\beta}_0 = -0.916 \) and \( \hat{\beta}_1 = 0.481 \).

Estimate of \( \sigma^2 \): ‘residual mean square’ = 3.170

c. First we compute \( r^2 = 1 - \frac{SS(\text{errors})}{SS(\text{Total})} \). We will use the ANOVA table (but one could also determine the sample variance of the \( y \)-values and use \( SS(\text{Total}) = (n - 1)S_y^2 \))

\[ r^2 = 1 - \frac{44.384}{116.000} = 0.6174 \], since \( \hat{\beta}_1 \) is positive, then so is \( r: r = \sqrt{0.6174} = 0.786 \)

d. 95%-Cl(\( \beta_1 \)) = \( \left( \hat{\beta}_1 - c \frac{S}{\sqrt{\Sigma(x_i-\bar{x})^2}}, \hat{\beta}_1 + c \frac{S}{\sqrt{\Sigma(x_i-\bar{x})^2}} \right) \), where \( P(T_{n-2} \geq c) = \frac{1}{2} \alpha \) (Property 3.1.5)

From the output we find \( \hat{\beta}_1 = 0.481 \) and the standard error \( \frac{S}{\sqrt{\Sigma(x_i-\bar{x})^2}} = 0.101 \) and from the \( t_{16-2} \)-table:

\( c = 2.145 \), so 95%-Cl(\( \beta_1 \)) = \( (0.481 - 2.145 \times 0.101, 0.481 + 2.145 \times 0.101) = (0.264, 0.698) \)

e. 1. \( y = \beta_0 + \beta_1 x + \epsilon \), where \( y \) = “yearly amount spend on food” and \( x \) = “the yearly income” and the independent disturbances \( \epsilon \) all have a \( N(0, \sigma^2) \)-distribution.

2. Test \( H_0: \beta_1 = 0 \) against \( H_1: \beta_1 \neq 0 \) with \( \alpha = 1\% \)

3. Test statistic: \( T = \frac{\hat{\beta}_1}{\sqrt{(\frac{S}{\Sigma(x_i-\bar{x})^2)}} \)

4. Under \( H_0: T \sim t_{n-2} = t_{14} \)

5. Observed value of \( T: \frac{0.481}{0.101} = 4.76 \) (output difference by rounding)

6. We will reject \( H_0 \) if \( T \leq -c \) or \( T \geq c \) at significance level 1%, so \( c = 2.977, \) such that \( P(T_{14} \geq c) = 0.005 \)

7. 4.76 lies in the Rejection Region \( \Rightarrow \) reject \( H_0 \)

8. We consider it proven, at a 1% level of significance, that the yearly income is related to the yearly amount spend on food.

Exercise 2
a. \( y = \beta_0 + \beta_1 x + \epsilon \), where \( y \) = “Turnover in Spring + Summer” and \( x \) = “Turnover in Autumn + Winter” and the independent disturbances \( \epsilon \) all have a \( N(0, \sigma^2) \)-distribution (the parameters \( \beta_0 \), \( \beta_1 \) and \( \sigma^2 \) are unknown parameters)

b. Just using the output: the estimates \( \hat{\beta}_0 \) and \( \hat{\beta}_1 \), respectively -11.489 and 0.98737, giving the equation of the estimated regression line: \( y = -11.489 + 0.98737x \).

c. The proportion explained variation is: \( R^2 = 1 - \frac{SS(\text{Errors})}{SS(\text{Total})} = 1 - \frac{7670.40}{91805.34} = 0.916% \).

This proportion is close to 1, so that the regression model is helpful to predict the turnover in Summer+ Spring based on the turnover in Autumn + Winter.

d. Prediction: \( \hat{y} = -11.489 + 0.98737 \times 140 = 126.7 \)

e. 95%-P(\( Y|x^* = 140 \)) has interval boundaries \( \hat{\beta}_0 + \hat{\beta}_1 x^* \pm c \sqrt{S^2 \left[ \frac{1}{n} + \frac{(x^*-\bar{x})^2}{\Sigma(x_i-\bar{x})^2} \right]} \),

where \( c = 2.160 \) is such that \( P(T_{15-2} \geq c) = \frac{1}{2} \alpha = 2.5\% \), \( \hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x^* = 126.7 \) (see part d.)

\( S^2 = 590.03 \) (\( MS(\text{Errors}) \) in ANOVA-table), \( n = 15, \bar{x} = 156.88 \) (given) and \( \Sigma(x_i-\bar{x})^2 = (n - 1)s^2_x = 14 \cdot (78.51)^2 = 86393.5 \):

the interval is \( (126.7 - 2.160 \times 25.13, 126.7 + 2.160 \times 25.13) = (72.4, 181.0) \)

( Compare to the confidence interval for the expected turnover in Spring/Summer: The 95%-Cl(\( E(Y|x^* = 140) \)) has interval boundaries \( \hat{\beta}_0 + \hat{\beta}_1 x^* \pm c \sqrt{S^2 \left[ \frac{1}{n} + \frac{(x^*-\bar{x})^2}{\Sigma(x_i-\bar{x})^2} \right]} \), where: \( x^* = 140 \)
and all other “ingredients” are the same as in part e. (the standard error is smaller now: 6.425): 95%-CI($\hat{\beta}_0 + \hat{\beta}_1 x^*$) = (126.7 − 2,160 · 6.425, 126.7 + 2,160 · 6.425) ≈ (112.8, 140.6).

f. The value 140.62 is pretty much in the middle of the prediction interval, established in e. Meaning that we cannot be sure that 140.62 is sufficient: if you would order 181 hl we can be 95% confident it is sufficient. Note that 140.6 hl is the upper bound of the 95% confidence interval, but this is an interval for the mean turnover (mean over many years) and does not give information about the turnover in a specific year.

Exercise 3

a. Estimates: $\hat{\beta}_0 = 162.281, \hat{\beta}_1 = -81.304$.

b. The standard errors: $se(\hat{\beta}_0) = 8.963$ and $se(\hat{\beta}_1) = 7.305$.

These are the estimates of the standard deviations of the underlying distributions of the estimates $\hat{\beta}_0$ and $\hat{\beta}_1$. We have $var(\hat{\beta}_1) = \sigma^2/S_{xx}$ with $S_{xx} = \sum(x_i - \bar{x})^2$

The standard deviation of $\hat{\beta}_1$ is thus $SD(\hat{\beta}_1) = \frac{\sigma}{\sqrt{S_{xx}}}$ estimated by the standard error $se(\hat{\beta}_1) = \frac{s}{\sqrt{n}} = 7.305$.

c. We are estimating $\sigma^2$ by means of $S^2 = \frac{\sum(y_i - \hat{\beta}_0 \hat{\beta}_1 x_i)^2}{n-2}$: we have to divide the residual sum of squares (numerator) by the corresponding degrees of freedom ($n - 2$) : we get 9.125

d. 95% confidence interval for $\beta_1$: boundaries are $\hat{\beta}_1 \pm c \times se(\hat{\beta}_1)$

Elaboration: $\hat{\beta}_1 = -81.304$ and $se(\hat{\beta}_1) = 7.305$ (from the output) $c = 2.306$ ($t$-distribution, $n - 2 = 8$ degrees of freedom, 95% probability between $-c$ and $c$)

95% confidence interval becomes: ($-81.304 - 2.306 \times 7.305, -81.304 + 2.306 \times 7.305) = (-98.1, -64.5$)

e. Let $Y$ denote the hardness and let denote $x$ the corresponding annealing temperature. The eight steps of the test:

1. $Y = \beta_0 + \beta_1 x + \epsilon$, with independent disturbances $\epsilon$ being $N(0, \sigma^2)$-distributed.

2. We test $H_0: \beta_1 = 0$ against $H_1: \beta_1 \neq 0$

3. Test statistic: $T = \frac{\hat{\beta}_1}{se(\hat{\beta}_1)}$

4. Under $H_0: T \sim t_{n-2} = t_8$

5. Outcome of $T$: $T = \frac{-81.304}{7.305} = -11.130$ (from output)

6. We reject $H_0$ if $T \leq -c$ or . $\alpha = 5\%$, table $\Rightarrow c = 2.306$

7. Rejection region contains outcome $-11.130$ so reject $H_0$.

8. Using level of significance 5% we have proven that the hardness depends on the annealing temperature.

f. Looking at the scatter plot we should see only ‘chaos’, no pattern whatsoever. We have to learn to judge residual plots. At first glance there is no pattern, perhaps caused by the fact that there are only a few points. Having a closer look you can distinguish a banana-shaped or V-shaped cloud of points vaguely. This has its origin in the plot of $y$ versus $x$: a slight curvature of the cloud of points can be distinguished as soon as you draw a straight line in the middle of the points. So a close examination of the plots may give rise to some doubt about the fit of the model. We return to this data set in another exercise. Then we investigate whether quadratic regression fits the data better.

Exercise 4

a. The model equation becomes: $Y = \beta_1 x + \epsilon$ (‘regression through the origin’)

We estimate the parameter $\beta_1$ by means of least squares:

Minimize $\sum(y_i - \beta_1 x_i)^2$ as function of $\beta_1$.

Differentiate: $-2 \sum(y_i - \beta_1 x_i) x_i$

Equating the derivative to zero gives an equation for the estimate $\hat{\beta}_1$: $-2 \sum(y_i - \beta_1 x_i) x_i = 0$

Solving: $\sum(y_i - \hat{\beta}_1 x_i) x_i = \sum x_i y_i - \hat{\beta}_1 \sum x_i^2 = 0$ and hence $\hat{\beta}_1 = \frac{\sum x_i y_i}{\sum x_i^2}$.

You can check that indeed a minimum has been attained: the second derivative is $+\sum x_i^2 > 0$.
b. Now we have to study the new estimator \( \hat{\beta}_1 = \sum x_i Y_i / \sum x_i^2 \) with independent random variables \( Y_i \) distributed according to a normal distribution with expectation \( \beta_1 x_i \) and (common) variance \( \sigma^2 \).

We compute its variance: 
\[
\text{var}(\hat{\beta}_1) = \text{var}(\sum x_i Y_i / \sum x_i^2) = \left( \sum x_i^2 \right)^{-1} \text{var}(Y_i) / \left( \sum x_i^2 \right)^2
\]

In general \( \sum x_i^2 > \sum (x_i - \bar{x})^2 = \sum x_i^2 - n \bar{x}^2 \) so the “new” variance of \( \hat{\beta}_1 \) tends to be smaller than the “old” variance \( \sigma^2 / \sum (x_i - \bar{x})^2 \).

**Exercise 5**

a. Let \( Y \) denote the mortality rate per 100 000 males. Let us first investigate whether the variable *North* really affects the mortality rate \( Y \), in addition to the variable *Calcium*. The eight steps of the test:

1. \( Y = \beta_0 + \beta_1 \text{Calcium} + \beta_2 \text{North} + \varepsilon \), with independent disturbances \( \varepsilon \) which are \( N(0, \sigma^2) \)-distributed.
2. We test \( H_0 : \beta_2 = 0 \) against \( H_1 : \beta_2 \neq 0 \)
3. Test statistic: \( T = \hat{\beta}_2 / se(\hat{\beta}_2) \)
4. Under \( H_0 : T \sim t_{58} \)
5. Outcome of \( T : 176.711 / 36.891 = 4.79 \)
6. We reject \( H_0 \) if \( T \leq c \) or \( T \geq c \). \( \alpha = 5\% \), \( t_{58} \Rightarrow c = 2.00 \) (interpolation)
7. The rejection region contains the outcome 4.79 so reject \( H_0 \).
8. Using level of significance \( 5\% \) we have proven that the variable *North* really affects the mortality rate \( Y \), in addition to the variable *Calcium*.

The test for investigating whether *Calcium* really affects mortality rate \( Y \), in addition to *North*, is as follows:

1. \( Y = \beta_0 + \beta_1 \text{Calcium} + \beta_2 \text{North} + \varepsilon \), with independent disturbances \( \varepsilon \) which are \( N(0, \sigma^2) \)-distributed.
2. We test \( H_0 : \beta_1 = 0 \) against \( H_1 : \beta_1 \neq 0 \)
3. Test statistics: \( T = \hat{\beta}_1 / se(\hat{\beta}_1) \)
4. Under \( H_0 : T \sim t_{61-3} = t_{58} \)
5. Outcome of \( T : -2.034 / 0.483 = -4.21 \)
6. We reject \( H_0 \) if \( T \leq c \) or \( T \geq c \). \( \alpha = 5\% \), \( t_{58} \Rightarrow c = 2.00 \) (interpolation)
7. The Rejection region contains the outcome -4.21 so reject \( H_0 \).
8. Using level of significance \( 5\% \) we have proven that the variable *Calcium* really affects the mortality rate \( Y \), in addition to the variable *North*.

b. We shall test whether it is useful to use the predictors *Calcium* and *North* for prediction of \( Y \). (According to some textbooks this should be the first step in a statistical analysis.) The eight steps of the test:

1. \( Y = \beta_0 + \beta_1 \text{Calcium} + \beta_2 \text{North} + \varepsilon \), with independent disturbances \( \varepsilon \) which are \( N(0, \sigma^2) \)-distributed.
2. We test \( H_0 : \beta_1 = \beta_2 = 0 \) against \( H_1 : \beta_1 \neq 0 \vee \beta_2 \neq 0 \)
3. Test statistic: \( F = \frac{SS(\text{regression})/k}{SS(\text{error})/(n-k-1)} = \frac{SS(\text{regression})/2}{SS(\text{error})/58} \)
4. Under \( H_0 : F \sim F_{58}^2 \)
5. Outcome of \( F : 1248317.8 / 864855.86 / 58 = 624158.905 / 14911.308 = 41.86 \)
6. We reject \( H_0 \) if \( F \geq c \). \( \alpha = 5\% \), \( F_{58}^2 \Rightarrow c = \frac{2}{20} \times 3.23 + \frac{18}{20} \times 3.15 = 3.16 \) (interpolation between \( F_{40}^2 \) and \( F_{60}^2 \))
7. Rejection Region contains outcome 41.86 so reject \( H_0 \).
We are searching for a pattern in the plot. There is no pattern, only chaos, this OK. No doubt about the fit.
Note we are looking in the vertical direction, judging the residuals. The points are not evenly spread in the horizontal direction. This makes the judgement about the residuals more difficult.
Furthermore we can search for outliers. Nearly all standardized residuals are contained by the interval $(-2,2)$. On the average only 5% of the standardized residuals should be lying outside $(-1.96, 1.96) \approx (-2,2)$. To our opinion the largest residual is not extreme enough to worry about.

This means that 57.7% of the spread of the dependent variable is explained by the two predictor variables. This is not a good value, weak relationship, if we want predict the dependent variable because it is far away from the optimal value 100%. But this time it is natural: It is a bad situation if you could predict the mortality rate perfectly by means of only the calcium concentration and geographic variable North.

Exercise 6

a. The eight steps of the test:

1. $Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$, with independent disturbances $\epsilon$ which are $N(0, \sigma^2)$-distributed.
2. We test $H_0: \beta_2 = 0$ against $H_1: \beta_2 \neq 0$
3. Test statistic: $T = \hat{\beta}_2 / \text{se}(\hat{\beta}_2)$
4. Under $H_0: T \sim t_{10-3} = t_7$
5. Outcome of $T: \frac{131.898}{40.648} = 3.25$
6. We reject $H_0$ if $T \leq -c$ or $T \geq c$, where $\alpha = 5 \% \Rightarrow c = 2.365$
7. The rejection region contains outcome 3.25 so reject $H_0$.
8. Using level of significance 5% we have proven that the quadratic term has to be added to the model.

We thus have proven that quadratic regression is better than simple linear regression.

b. We have to judge whether the residual plot shows chaos. That seems to be the case. It is a little bit troublesome that there are more points in the left part of the plot, suggesting unequal spread of the residuals. I guess we need more data for assessing such a pattern.

c. Modified scheme of eight steps:

1. $Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$, with independent disturbances $\epsilon$ which are $N(0, \sigma^2)$-distributed.
2. We test $H_0: \beta_2 = 0$ against $H_1: \beta_2 \neq 0$
3. Test statistic: $T = \hat{\beta}_2 / \text{se}(\hat{\beta}_2)$
4. Under $H_0: T \sim t_{10-3} = t_7$
5. Outcome of $T: \frac{131.898}{40.648} = 3.25$
6. The (two-sided) p-value is 0.014
7. Since p-value $\leq \alpha = 5 \%$ we have to reject the null hypothesis.
8. Using level of significance 5% we have proven that the quadratic term has to be added to the model.

We thus have proven that quadratic regression is better than simple linear regression.

If we would choose $\alpha = 1 \%$ then p-value= 0.014 $> \alpha$, and hence we don’t reject $H_0$.

Exercise 7

a. For answering the question we have to rearrange the model equation:

$Y = (\beta_0 + \beta_2 x_2) + (\beta_1 + \beta_3 x_2) \times x_1 + \epsilon$

If we fix $x_2$ then $Y$ depends on $x_1$ in a linear way, $\beta_0 + \beta_2 x_2$ is the (new) constant (intercept) and $\beta_1 + \beta_3 x_2$ is the (new) slope. Note that this slope is affected by $x_2$, this disappears when we skip the interaction term.

b. We copy the output and replace all signs “?”.
c. The estimated equation: \( y = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_1 x_2 \)

We get from the output: \( y = 14.9600 + 1.5321 x_1 - 0.4323 x_2 - 0.0553 x_1 x_2 \)

We calculate a (point)prediction by using the estimated equation with \( x_1 = 12 \) and \( x_2 = 3 \). This renders the following prediction: \( \hat{y} = 14.9600 + 1.5321 \times 12 - 0.4323 \times 3 - 0.0553 \times 12 \times 3 = 30.06 \)

d. Prediction for \( x_1 = 12 \) en \( x_2 = 4 \): \( \hat{y} = 14.9600 + 1.5321 \times 12 - 0.4323 \times 4 - 0.0553 \times 12 \times 4 = 28.96 \)

This is strange because the last prediction is smaller than the first prediction.

We should think about the interpretation of the individual parameters \( \beta_i \):

The parameter \( \beta_i \) in general reflects the mean change in the dependent variable \( Y \) when we increase \( x_i \) with 1 unit and fix the other predictor variables.

Fixing the other predictor variables while changing one specific predictor variable does not reflect reality often, because many times predictor variables are changing simultaneously:

Here raising advertisement costs and increasing the number of sales representatives may be simultaneous actions. Then distinguishing the respective causes of the predictor variables may be difficult because of the dependence the predictors.

In statistics it is a famous phenomenon that estimates \( \hat{\beta}_i \) turn out to have the wrong sign (you expected a positive value but you get a negative value or vice versa). Many times this phenomenon can be explained by strong relationships between predictor variables, to some extent a number of predictor variables share the same information. Sometimes the ‘strangeness’ can be solved by simplifying the model.

e. Boundaries for 95% confidence interval for \( \beta_2 : \beta_2 \pm c \times se(\hat{\beta}_2) \),

with -0.4323 and 1.7964 for \( \hat{\beta}_2 \) and \( se(\hat{\beta}_2) \) respectively, and \( c = 2.31 \) (\( t_8 \)-distribution).

We get: \((-0.4323 - 2.31 \times 1.7964, -0.4323 + 2.31 \times 1.7964) = (-4.5823, 3.717) \)

f. The eight steps of the test:

1. Model: \( Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 \times x_2 + \varepsilon \), with independent disturbances \( \varepsilon \) which are \( N(\mu, \sigma^2) \)-dist.
2. We test \( H_0 : \beta_3 = 0 \) against \( H_1 : \beta_3 \neq 0 \)
3. Test statistic: \( T = \hat{\beta}_3 / se(\hat{\beta}_3) \)
4. Under \( H_0 \) we have \( T \sim t_8 \)
5. Observed value \( t = \frac{-0.0553}{0.1594} \approx -0.36 \)
6. We reject \( H_0 \) if \( T \leq -c \) or \( T \geq c \). Level of significance 5%, \( t \)-table \( \Rightarrow c = 2.31 \)
7. The rejection region does not contain -0.36, hence we don’t reject \( H_0 \).
8. Using level of significance 5% we did not prove that the interaction term should be part of the model.

When you apply again the t-test for testing \( H_0 : \beta_2 = 0 \) against \( H_0 ; \beta_2 \neq 0 \) then again you need not reject the null hypothesis. This does not mean that you have to skip both terms \( \beta_2 x_2 \) and \( \beta_3 x_1 \times x_2 \). Skip first the interaction term (a model with the interaction term and without the term \( \beta_2 x_2 \) is weird). Continue with the model without interaction and test \( H_0 : \beta_2 = 0 \) against \( H_0 ; \beta_2 \neq 0 \) in order to investigate whether we can skip \( x_2 \) completely.

For this procedure we thus need more output.

h. Scatter plots of the data is always helpful for appreciating the model applied. Furthermore a model check by means of a scatter plot of residuals is missing.

Exercise 10 (Exercise 8 and 9 cancelled)

a. 1. Model: \( Y = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon, i = 1,2,3 \) and \( j = 1,2,3 \), where:

the \( \alpha_i \)'s are the levels of the factor Material and the \( \beta_j \)'s are the levels of the factor Temperature

\( \sum_{i=1}^{3} \alpha_i = 0 \) , \( \sum_{j=1}^{3} \beta_j = 0 \) , \( \sum_{i=1}^{3} \gamma_{ij} = 0 \) for each \( j \) and \( \sum_{j=1}^{3} \gamma_{ij} = 0 \) for each \( i \)

and for the observations the errors (disturbances) \( \varepsilon \) are assumed independent and all \( N(0, \sigma^2) \).

The results of analysis of variance by SPSS (the 3 columns of the SPSS-file are given in the first table under b.)
2. The p-value of the test on \( H_0: \alpha_i = \beta_j = \gamma_{ij} = 0 \) for \( i, j = 1, 2, 3 \) against \( H_1: \text{at least one coefficient } \neq 0 \) has a p-value \( PF \geq 11.000 (P_H) \approx 0.000 < 0.0005 < \alpha = 5\% \). Hence we can conclude (at 5\% level of significance) that at least one of the effects of material, temperature or interaction is significant.

3. The p-value of the test on \( H_0: \gamma_{ij} = 0 \) for \( i, j = 1, 2, 3 \) against \( H_1: \gamma_{ij} \neq 0 \) for at least one pair \((i, j)\), in the model with the factors Material and Temperature, has a p-value \( 1.9\% < \alpha = 5\% \). Hence we can conclude (at 5\% level of significance) that adding the interaction effects to the model is beneficiary in explaining the life of batteries.

4. The p-value of the test on \( H_0: \alpha_i = 0 \) for \( i = 1, 2, 3 \) against \( H_1: \alpha_i \neq 0 \) for at least one value of \( i \), in the model with the other factors Temperature and Interaction, has a p-value \( 0.2\% < \alpha = 5\% \). Hence we can conclude (at 5\% level of significance) that adding the Material factor to the model increases the explanation of the life of batteries significantly.

5. The p-value of the test on \( H_0: \beta_j = 0 \) for \( j = 1, 2, 3 \) against \( H_1: \beta_j \neq 0 \) for at least one value of \( j \), in the model with the other factors Material and interaction, has a p-value \( 1.9\% < \alpha = 5\% \). Hence we can conclude (at 5\% level of significance) that adding the temperature factor to the model enhances the explanation of the life of batteries significantly.

b. Below you see first the data matrix (in SPSS), then first the results of the linear regression with all (8) variables and (extra!) then without the (4) interaction terms.
Running Linear regression with all 8 variables (including interaction):

<table>
<thead>
<tr>
<th>Material 2</th>
<th>Material 3</th>
<th>Medium Temp</th>
<th>High Temp</th>
<th>x1*x3</th>
<th>x1*x4</th>
<th>x2*x3</th>
<th>x2*x4</th>
</tr>
</thead>
<tbody>
<tr>
<td>122</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>106</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>115</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>25</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>70</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>58</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>45</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>138</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>110</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>168</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>160</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>174</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>120</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>150</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>139</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>96</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>104</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>82</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>60</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

**Model Summary**

<table>
<thead>
<tr>
<th>Model</th>
<th>R</th>
<th>R Square</th>
<th>Adjusted R Square</th>
<th>Std. Error of the Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.875</td>
<td>.765</td>
<td>.696</td>
<td>25.985</td>
</tr>
</tbody>
</table>

a. Predictors: (Constant), x2*x4, x2*x3, x1*x4, x1*x3, High temp, material 2, material 3, Medium temp

**Coefficients**

<table>
<thead>
<tr>
<th>Model</th>
<th>(Constant)</th>
<th>Material 2</th>
<th>Material 3</th>
<th>Medium Temp</th>
<th>High Temp</th>
<th>x1*x3</th>
<th>x1*x4</th>
<th>x2*x3</th>
<th>x2*x4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>134,750</td>
<td>21,000</td>
<td>9,250</td>
<td>-77,500</td>
<td>-77,250</td>
<td>-29,000</td>
<td>41,500</td>
<td>79,250</td>
<td>18,750</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model</th>
<th>B</th>
<th>Std. Error</th>
<th>Beta</th>
<th>t</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>134,750</td>
<td>12,992</td>
<td>.213</td>
<td>10,371</td>
<td>.000</td>
</tr>
<tr>
<td></td>
<td>21,000</td>
<td>18,374</td>
<td>.094</td>
<td>1,143</td>
<td>.263</td>
</tr>
<tr>
<td></td>
<td>9,250</td>
<td>18,374</td>
<td>-.787</td>
<td>-4,218</td>
<td>.000</td>
</tr>
<tr>
<td></td>
<td>-77,500</td>
<td>18,374</td>
<td>-.784</td>
<td>-4,204</td>
<td>.000</td>
</tr>
<tr>
<td></td>
<td>-77,250</td>
<td>18,374</td>
<td>-.196</td>
<td>-1,116</td>
<td>.274</td>
</tr>
<tr>
<td></td>
<td>-29,000</td>
<td>25,985</td>
<td>.536</td>
<td>3.050</td>
<td>.005</td>
</tr>
<tr>
<td></td>
<td>41,500</td>
<td>25,985</td>
<td>.127</td>
<td>.722</td>
<td>.477</td>
</tr>
<tr>
<td></td>
<td>79,250</td>
<td>25,985</td>
<td>.356</td>
<td>3.050</td>
<td>.005</td>
</tr>
<tr>
<td></td>
<td>18,750</td>
<td>25,985</td>
<td>.127</td>
<td>.722</td>
<td>.477</td>
</tr>
</tbody>
</table>

a. Dependent Variable: Life (in hours)
Results, running linear regression in SPSS with 4 variables (without interaction):

<table>
<thead>
<tr>
<th>Model</th>
<th>Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Regression</td>
<td>59416,222</td>
<td>8</td>
<td>7427,028</td>
<td>11,000</td>
</tr>
<tr>
<td></td>
<td>Residual</td>
<td>18230,750</td>
<td>27</td>
<td>675,213</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>77646,972</td>
<td>35</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

a. Predictors: (Constant), \( x_2 \times x_4, x_2 \times x_3, x_1 \times x_4, x_1 \times x_3 \), High temp, material 2, material 3, Medium temp

b. Dependent Variable: Life (in hours)

1. The F-test in the ANOVA table (all 8 variables) reports the value 11.000 for \( F = \frac{MS(Regression)}{MS(Errors)} \), with a p-value \( P(F \geq 11.000|H_0) < 0.0005 \), these values are the same as reported in a.1 (two-factor analysis).

2. Both the coefficients for the factor “Material” seem to be not significant (in the presence of the other in the model). We cannot conclude that Material a factor can be cancelled: there can be a strong relation between these variables (“collinearity”), such that one variable is not significant in the presence of the other in the model. This confirmed by the p-value of the test on the factor “Material” in a.3: the factor has significant effects. Possibly one of the variables can be cancelled: e.g. cancelling \( x_2 \) (material 3 has the largest p-value) would mean that we only distinguish material 2 and not-material 2. One can take the value of \( R^2_{adj} \) after removing \( x_2 \) in consideration.

Both the coefficients of the factor Temperature are significant (p-value = 0.000 < \( \alpha \)). 3 out of 4 interaction coefficients can be potentially be cancelled from the model since the p-value > \( \alpha = 5\% \).
3. Note that adding the interaction (from 4 to 8 variables) increased the $R_{adj}^2$ (from 59.5% to 69.6%), justifying this increase in number of variables: but canceling $x_2$ or one or more interaction variables, might increase the value. Check that if we (e.g.) cancel $x_0$ the value $R_{adj}^2$ increases to 70.1%.

4. Based on (only) the information of the “full” model with interaction, we cannot test on the factor “Interaction”, as described in a.3 of analysis of variance. But running Linear regression for the null-model without interaction as well we can (see property 3.2.8 and example 3.3.10):

$$F = \frac{[SS_0(\text{Errors}) - SS(\text{Errors})]}{SS(\text{Errors})} = \frac{(27844.528 - 18230.750)/4}{18230.750/27} \approx 3.56$$

(almost) the same value of $F$ we found in a.3: the RR is $F \geq 2.73$ ($\alpha = 5\%$, $F_{27}^4$-table), so reject the null hypothesis of “no effect of interaction”, as before.